DOCUMENT RESUME

ED 445 522 FL 026 394

AUTHOR	Tokmouline, Timur
TITLE	Artificial Intelligence in Computerized Linguistic Analysis.
PUB DATE	2000-00-00
NOTE	25p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Algorithms; Artificial Intelligence; *Computational Linguistics; *Computer Assisted Testing; Educational Technology; Interference (Language); Language Research; Linguistic Theory; *Structural Analysis (Linguistics); *Test Construction

ABSTRACT

This article explores whether or not it is possible for computers to be effectively used to analyze textual information. Computerization of human linguistic analysis would be enormously useful, because it would relieve many highly skilled linguistics professionals (researchers and teachers) from having to spend enormous amounts of time on the tedious task of analyzing the composition of tests based on educational materials. A general algorithm operationalized in a computer program is proposed for linguistic analysis. The researcher concludes that the proposed algorithm is effective, because it has been demonstrated to analyze a predicate of a simple informative sentence, effectively automating the task of linguistic analysis. It is asserted that this is a significant breakthrough in the field. Several matrices and screen captures appear throughout the article. (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 522

Artificial Intelligence in Computerized Linguistic Analysis

By Timur Tokmouline

Every day human beings read newspapers, books, and engage in conversations with others. All of these activities are forms of human communication and involve clever manipulation and interpretation of the obtained information in form of textual data. Textual data is classified according to the nature of its content: scientific, literary, public, conversational, etc. The process of manipulation and interpretation of textual data is better known as linguistic analysis or natural language understanding.

From the official beginning of the computer science in early 1950's, computer scientists have attempted to create an intelligent computer, which would be capable of analyzing textual data. Many resulting programs appeared to work well, however "under the hood," they were mere matching programs, trying to relate a supplied sentence with a sentence structure stored somewhere in the program database. The program then used this structure to compose questions according to found sentence structure. In other words, the analysis done by such program yielded nothing more than a clever manipulation of the input, whereas the human linguistic analysis could determine the nature of the action portrayed by the input. Thus, it is evident that to this day, the human linguistic analysis has not been simplified and formalized down to the computer algorithm.

Although the process of linguistic analysis is a vivid sign of intelligence, it is also a very boring and tedious task that occupies a lot of time. For example, school teachers as well as college professors spend enormous amounts of time on composition of tests based on educational materials. The same way, many educational agencies that make

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Timur Tokmouline

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

□ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL026394

standardized tests, such as Educational Testing Service, hire people to analyze textual information from many books for composition of those tests. Perhaps computerization of human linguistic analysis can relieve many people from spending long hours on nothing but text analysis; after all, time is money! But is this possible? In other words, can the human linguistic analysis be simplified and formalized down to a computer algorithm?

In order to be able to answer this question, we must first explore the basics of linguistics. First, it is a fact that all textual data is composed of fundamental text units called sentences. All of the sentences are known to be "extensions" of simple sentences. Hence, in order for one to be able to analyze textual data, he or she must first be able to analyze a simple sentence. The Science of Linguistics defines a simple sentence to be the most abstract type of sentence, having only one subject and only one predicate. It is also very well known that such sentences may be of different types: affirmative, negative, interrogative, imperative, informative, etc.

All textual data is classified according to the nature of its context: scientific, literary, public, conversational, etc. This experiment uses the analysis of English scientific textual data, because this kind of textual data is usually the main "supplier" of the information and follows a very formal set syntactic structure. Such textual data is largely made up of simple informative sentences (sentences that give new information to the reader). Hence, if it is possible to analyze a simple informative sentence, it is possible to analyze scientific textual data.

But what does "analyze a simple informative sentence" mean? In order to answer this question, it is necessary to consider the more important members of the sentence.

"A sentence consists of two parts: the subject and the predicate. The subject of the sentence is the part about which something is being said. The predicate is the part which says something about the subject"(Warriner, 40). From this we can infer that the nature of the action portrayed by the sentence, or the nature of the message carried by the sentence, is enclosed in the predicate. Hence to be able to analyze the predicate is to be able to analyze the sentence.

In order to analyze predicate, it is imperative to first single out all possible forms in which predicate exists. Linguistics always defined predicate as a verb or as a combination of verbs, and it is well known that verbs tell about time, aspect and voice. Actions in all languages take place either in the past, present, or future tenses. Moreover, each verb has its own aspect, or its own basic nature. An aspect of a verb can be indefinite, continuous, perfect, and perfect continuous. Verbs with indefinite aspect describe actions that take place once or constantly repeat. Verbs with continuous aspect depict actions that have no beginning or end. Those verbs that have perfect aspect portray actions that have no beginning but do have an end. The verbs that have perfect continuous aspect describe actions that have no beginning and no end, but yet do produce some sort of result. The verbs that describe the actions done directly by the subject are said to have active voice. Likewise, the verbs that describe the actions that are performed on the subject but not by the subject are said to have passive voice. Table 1 accurately shows all the possibilities of composition of predicate with active voice, and table 2 shows all the possibilities of composition of predicate with passive voice.

Linguistic studies conducted by scientists (linguists) beginning in 1960 resolved the more frequently used possibilities of predicate compositions into a geometric figure, a

ACTIVE

	Past	Present	Future
Indefinite	-ed;II	∅,-s,-es	Shall +infinitive Will
Continuous	Was +ing Were	Am Is +ing Are	Shall +be+ing Will
Perfect	Had +ed III	Have + ed Has III	Shall +been+ed Will III
Perfect Continuous	Had+ been+ing	Have +been+ing Has	Shall +have+been+ing Will

Table 1. Active Voice

BEST COPY AVAILABLE

ERIC
Full Text Provided by ERIC

PASSIVE

	Past	Present	Future
Indefinite Be+ed III	Was + ed Were III	Am Is +ed Are III	Shall +be+ed Will III
Continuous Be being ed III	Was +being+ed Were III	Am Is +being+ed Are III	Shall +be+being+ed Will III
Perfect Have+been+ed III	Had+been+ed III	Have +been+ed Has III	Shall +have+been+ed Will III
Perfect *Continuous* Have+been+being+ed III	Had+been being+ed III	Have +been+being+ed Has III	Shall +have+been+being+ed Will III

Table 2. The Passive Voice

parallelogram (See figure 1), where the tense, the aspect and the voice are better diagrammed.

All of the forms of predicates listed in Table 1 and in Table 2 are only recognized initially as possibilities; some forms of predicates (with the passive voice) are not used in daily language because they are too complex. Whether one of the candidate forms on either one of these tables is determined by many other factors which surround the verbs that compose a candidate form for the predicate.

Hence, to analyze the predicate, is to determine the composition (all the verbs that compose it), the tense, the aspect, and the voice of the predicate. Now, using the transitive property of logic(if a leads to b and b leads to c, then a leads to c), we can conclude that simplification and formalization of human linguistic analysis down to a computer algorithm is possible if and only if a thorough analysis (determination of tense , aspect, composition, and voice) of predicate of a single sentence is possible.

In order for any analysis to be done on the sentence, the sentence must first be broken down into words. Then, every word must be identified to a part of speech (noun, verb, adjective, etc.) So the first step in such algorithm is parsing of sentence (breaking down of the sentence into words and then recognition of individual words as parts of speech). The problem that arises right away is the ambiguity; in English, some words can serve as verbs, adjectives, and nouns in different situations. The first step to solve this problem is make a list of parts of speech for each word (e.g. for can, such list is "verb noun"). Then, we must then use some common lexical assumptions (listed below) to single out the part of speech to which each word belongs. English language presents a rather large number of possibilities of structures, thus making it nearly impossible to

BEST COPY AVAILABLE

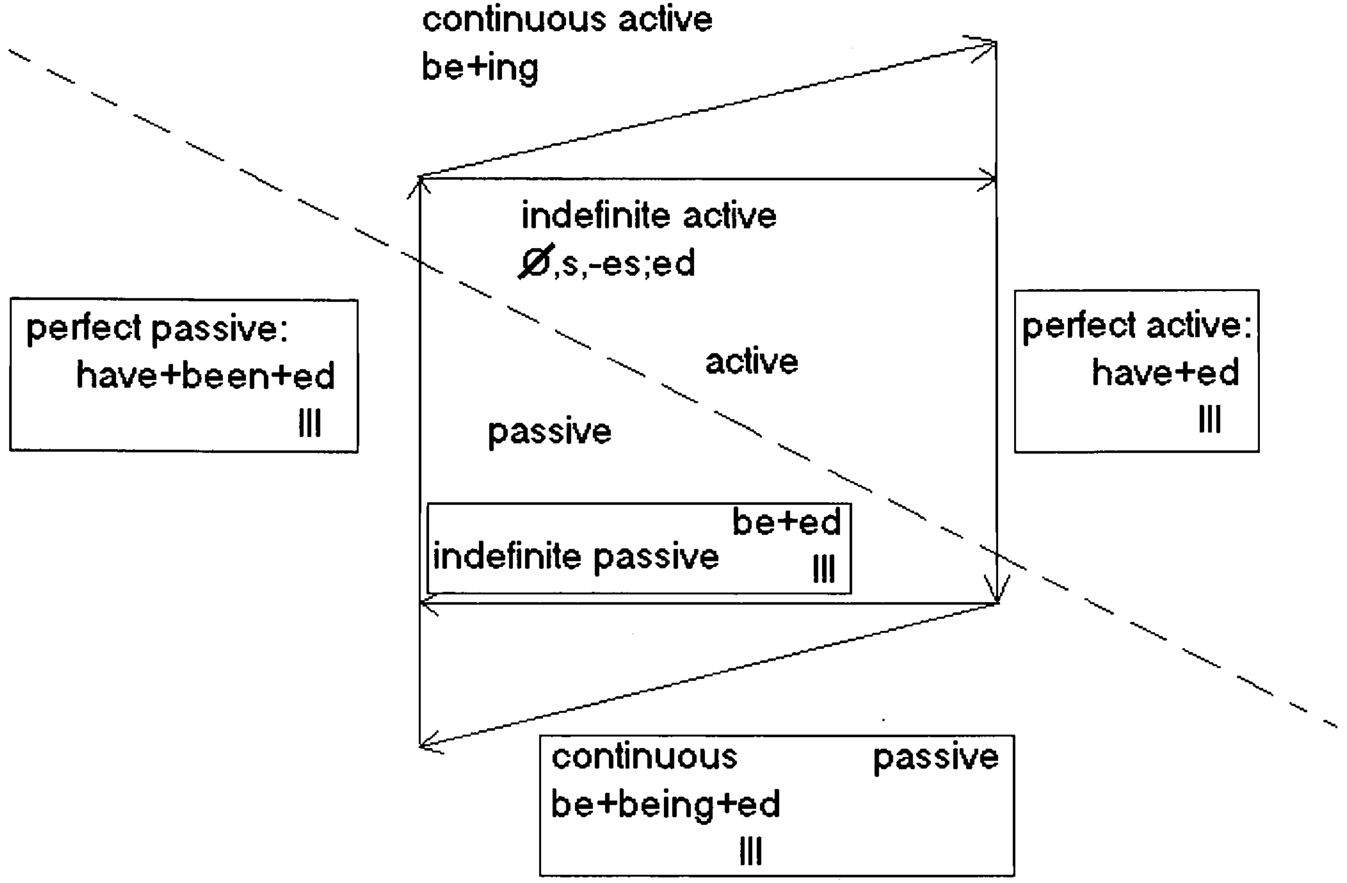


Figure 1 Everything above the dotted line has active voice. Everything below the dotted line has passive voice. Every possible composition of verbs that is a candidate for predicate stationed on the horizontal line has an indefinite aspect. Every form that is a candidate for predicate stationed on the vertical line has a perfect aspect. Every form that is a candidate for predicate stationed on the slanted line has a continuous aspect.

account for all possibilities. Therefore, the listed assumptions will not get rid of ambiguity one hundred percent of the time.

- The word that comes right after an article is a noun.
- The word that comes right after a preposition is either an article or a noun.
- The word that comes right after a noun is a verb.
- A word that comes after a modal verb is also a verb.
- A word that comes after a verb and seems to be a derivative of some verb with an “ing” ending is also a verb.
- A word that comes after a preposition and has an “ed” ending is an adjective. (If it doesn’t have an “ed” ending, it is a noun).
- A word that ends with “ed” and comes after "am", "is", "are", "was", "were", "will", "would", "shall", or "should" and before “to”, “for “, “from”, “in”, “on”, or “of,” is a verb.
- A word that ends in “ed” and comes after a verb and is a derivative of some verb, is also a verb.
- A word that ends in “ed” and comes before an article but not after a verb, is a verb.
- A word that ends in “ing” and is the first word of the sentence, is a noun.
- A word that ends in “ing” and comes before an article is a verb.

After the ambiguity problem is resolved, it is necessary to proceed to the core of the analysis: finding the possible candidates for a predicate of the sentence. The traditional approach to any analysis that traditional computer science takes involves a consideration of as many possibilities as possible. All candidates for a predicate in a sentence must be expressed in one of the possible candidate forms displayed either in Table 1 or in Table 2.

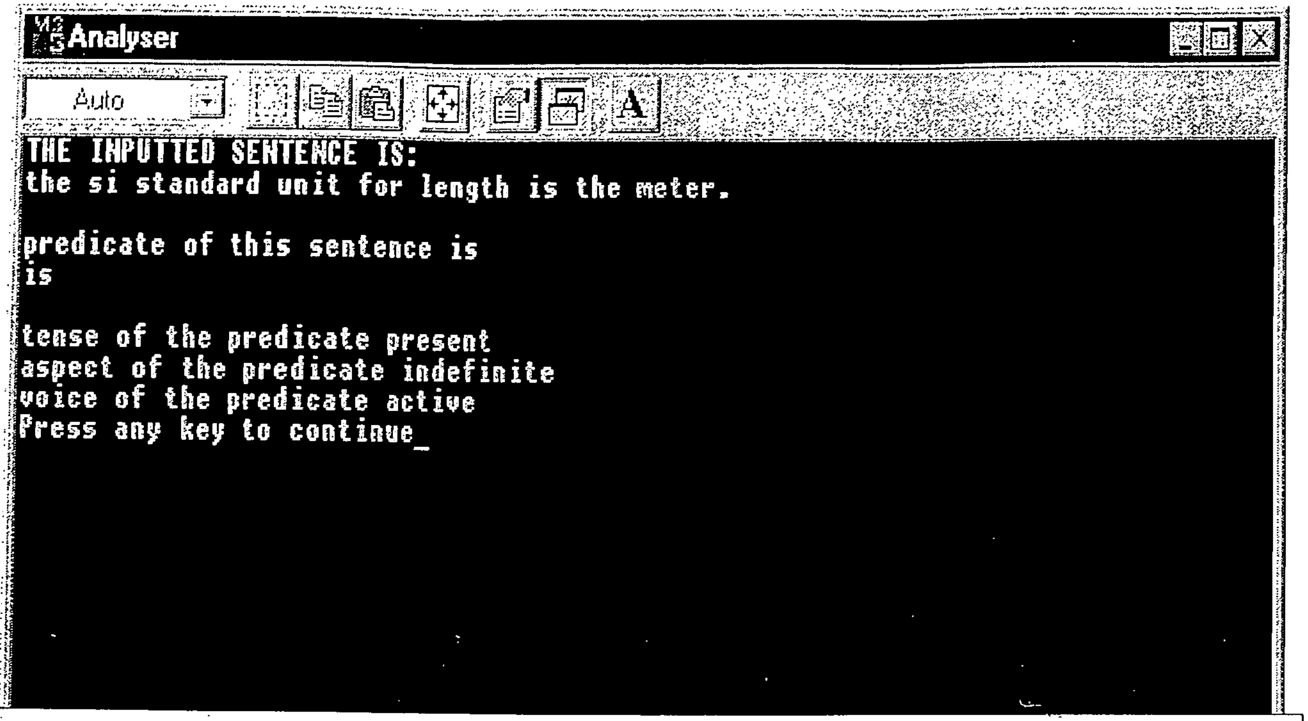
BEST COPY AVAILABLE

If a candidate satisfies one of the forms mentioned either in Table 1 or in Table 2, it is the predicate of the sentence.

Now that the general algorithm for finding the predicate in a simple, informative sentence has been developed in theory, there is one step left that is needed to be completed in order for the experimental problem to be thoroughly answered. That crucial step is the testing of the developed algorithm with hopes that it works. The best way to accomplish this (also the only way) is to build a program, which would be based on this algorithm. This is exactly what was done (by me), and the source code for that program is available upon request. So if one knows if the program works, then one also knows that the algorithm also works successfully. But how is it possible to know if the program works? One possible way to do this is to feed a set of simple sentences to the program, and then observe if the program would find the predicate in those sentences correctly. If the program does correctly identify the predicate, then the algorithm works successfully. Thus, the following sentences were fed to that program:

- The SI standard unit for length is the meter.
- The word atom is from the Greek word meaning indivisible.
- All matter is composed of extremely small particles called atoms.
- John Dalton was interested in the composition and properties of gases.
- Atomic structure refers to the identity and arrangement of smaller particles within atoms.
- Atomic theory has been expanded to explain the new observations.

The output of the analyzing program for each one of the sentences is attached.



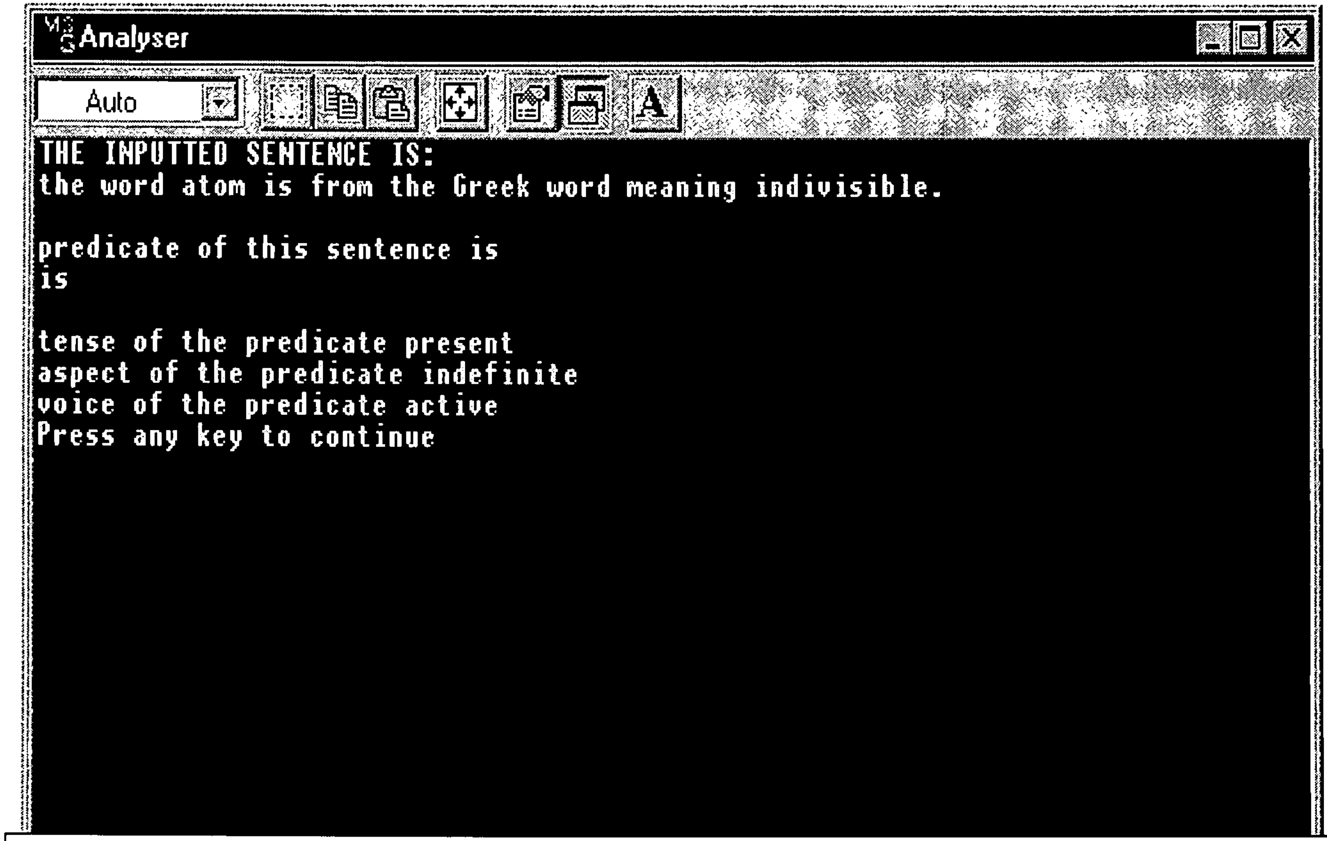
```
THE INPUTTED SENTENCE IS:
the si standard unit for length is the meter.

predicate of this sentence is
is

tense of the predicate present
aspect of the predicate indefinite
voice of the predicate active
Press any key to continue_
```

The predicate of "The SI standard unit for length is the meter" is the "is", which in this sentence has a present tense, indefinite aspect, and an active voice. As you can see, the program based on the algorithm correctly identified and analyzed the predicate.

BEST COPY AVAILABLE



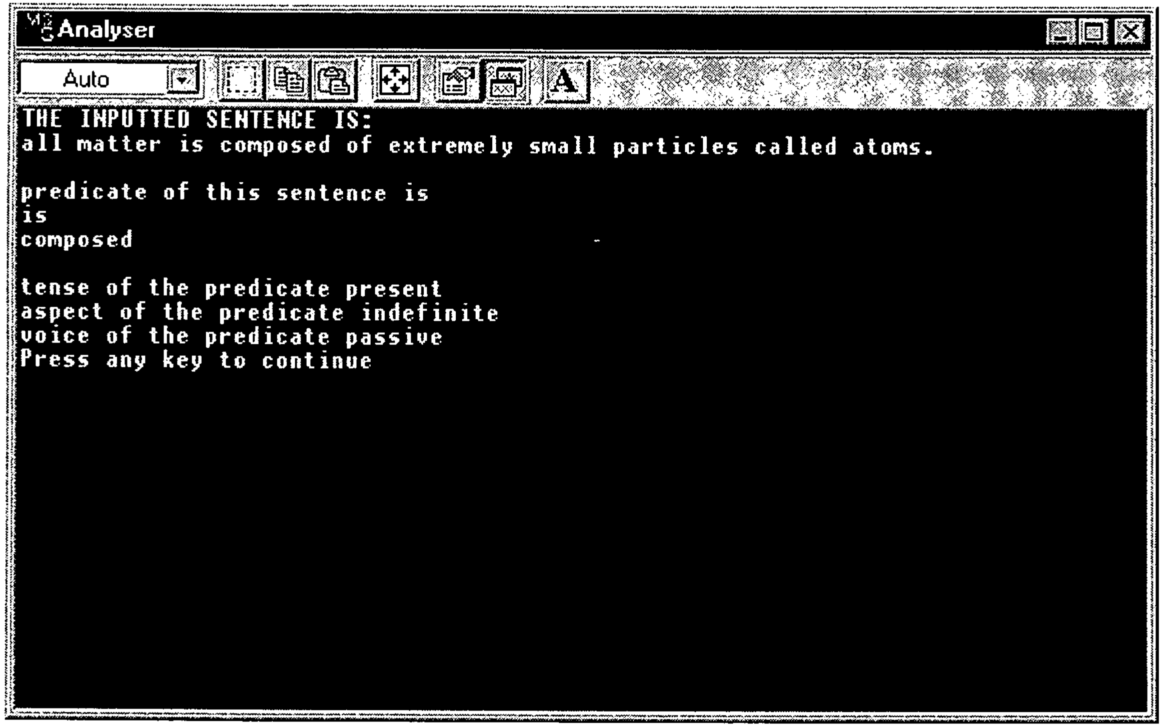
```
THE INPUTTED SENTENCE IS:
the word atom is from the Greek word meaning indivisible.

predicate of this sentence is
is

tense of the predicate present
aspect of the predicate indefinite
voice of the predicate active
Press any key to continue
```

The predicate of “The word atom is from Greek word meaning indivisible” is again “is” which in this sentence also has a present tense, indefinite aspect, and an active voice. As you can see, the program based on the algorithm correctly identified and analyzed the predicate.

BEST COPY AVAILABLE



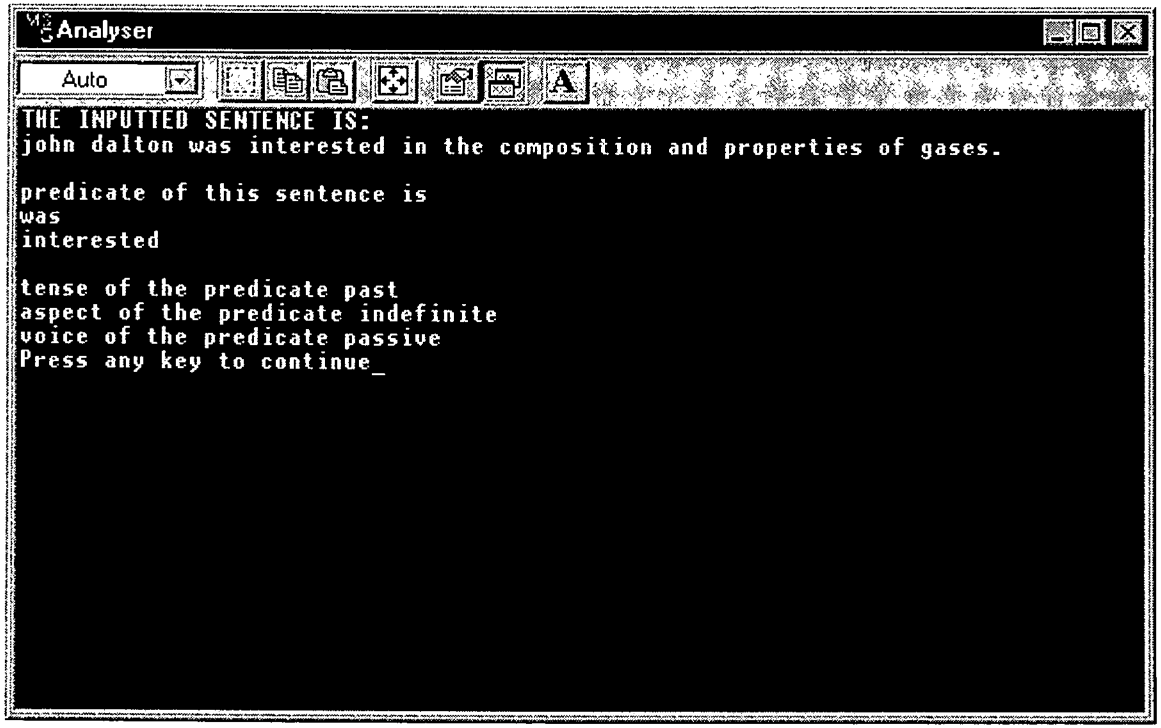
```
THE INPUTTED SENTENCE IS:
all matter is composed of extremely small particles called atoms.

predicate of this sentence is
is
composed

tense of the predicate present
aspect of the predicate indefinite
voice of the predicate passive
Press any key to continue
```

The predicate of "All matter is composed of extremely small particles called atoms" is "is composed" and in this sentence is in present tense, indefinite aspect, and an active voice. As you can see, the computer correctly analyzed and identified the predicate in this sentence.

BEST COPY AVAILABLE



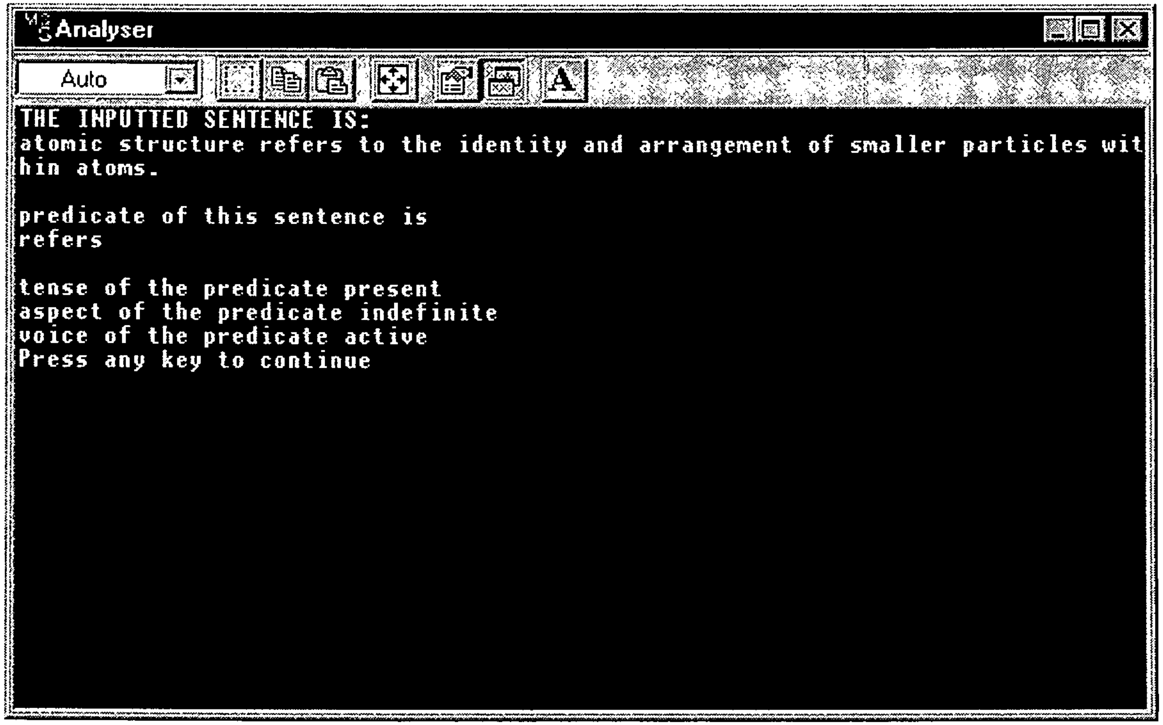
```
THE INPUTTED SENTENCE IS:
john dalton was interested in the composition and properties of gases.

predicate of this sentence is
was
interested

tense of the predicate past
aspect of the predicate indefinite
voice of the predicate passive
Press any key to continue_
```

The predicate of "John Dalton was interested in the composition and properties of gases" is "was interested" and in this sentence is in past tense, has an indefinite aspect, and a passive voice. As you can see, the computer correctly analyzed and identified the predicate in this sentence.

BEST COPY AVAILABLE



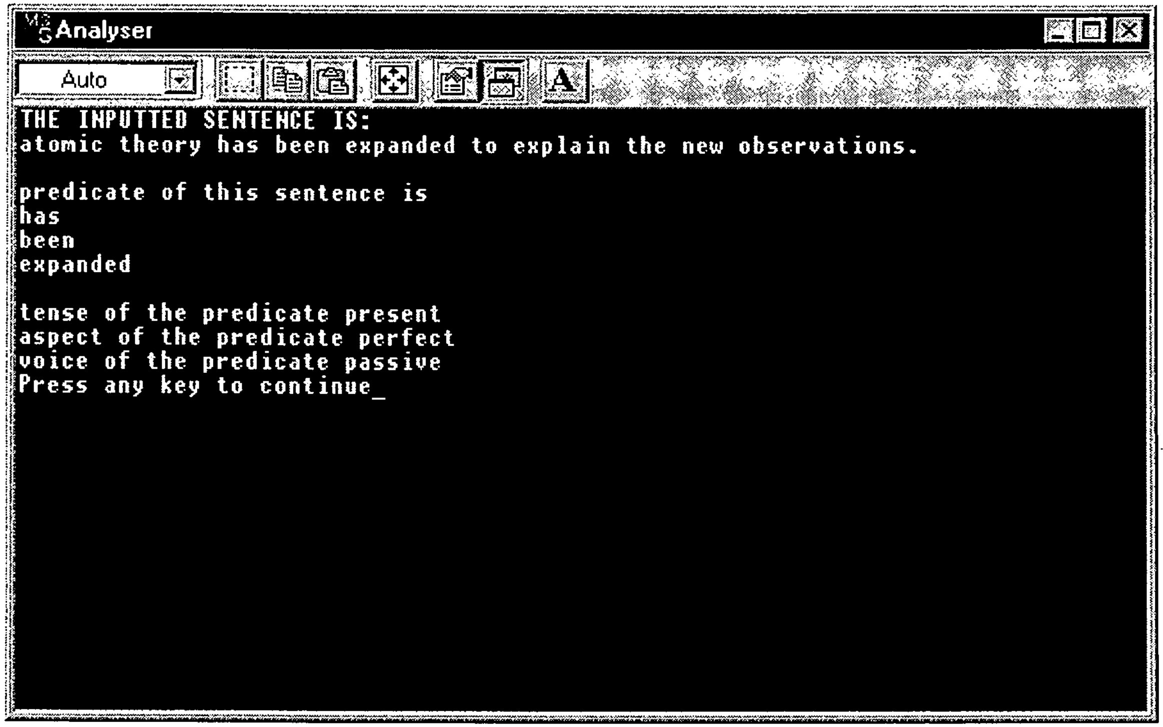
```
THE INPUTTED SENTENCE IS:
atomic structure refers to the identity and arrangement of smaller particles wit
hin atoms.

predicate of this sentence is
refers

tense of the predicate present
aspect of the predicate indefinite
voice of the predicate active
Press any key to continue
```

The predicate of “Atomic structure refers to the identity and arrangement of smaller particles within atoms” is “refers” and in this sentence is in present tense, indefinite aspect, and an active voice. As you can see, the computer correctly analyzed and identified the predicate in this sentence.

BEST COPY AVAILABLE



```
THE INPUTTED SENTENCE IS:
atomic theory has been expanded to explain the new observations.

predicate of this sentence is
has
been
expanded

tense of the predicate present
aspect of the predicate perfect
voice of the predicate passive
Press any key to continue_
```

The predicate of "Atomic Theory has been expanded to explain the new observations" is "has been expanded" and in this sentence is in present tense, continuous aspect, and a passive voice. As you can see, the computer correctly analyzed and identified the predicate in this sentence.

BEST COPY AVAILABLE

It is now evident that the program has correctly identified and analyzed the predicate of each one of the six supplied sentences, because each predicate was correctly located and its tense, aspect, and voice correctly determined. From this, one can directly conclude that the program and the algorithm perform as expected, yielding correct analysis. As you can see, the working algorithm that can successfully analyze a predicate of a simple informative sentence now does exist. Hence, because the algorithm that can successfully analyze a predicate of a simple informative sentence now does exist, the algorithm that is formalized and simplified version of human linguistic analysis also does exist. The amount of possibilities that needed to be considered to merely find a predicate of a simple informative sentence is massive. From this we can infer that the amount of possibilities that will need to be considered is absolutely prodigious.

Bibliography

1. Allen, James. Natural Language Understanding. Benjamin/Cummings Publishing company, Inc 1995.
2. Copeland, Jack. Artificial Intelligence. Blackwell publishers 1993.
3. Fetzer, James H. Artificial Intelligence and Its Scope and Limits. Kluwer Academic Publishers, Boston 1990.
4. Joufoni, Graciella. Communication Patterns and Textual Forms. Cromwell Press, Wiltshire 1996.
5. Warriner, John E.; Whitten, Mary E; Griffith, Francis. Warriner's English Grammar and Composition, third course. Heritage ed. New York: Harcourt Brace Jovanovich, 1977.

FL020394

ERIC

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Artificial Intelligence in Computerized Linguistic Analysis	
Author(s): Timur Tokmouline	
Corporate Source:	Publication Date: 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here please

Signature: Timur Tokmouline	Printed Name/Position/Title: Timur Tokmouline	
Organization/Address: 67 Rita Dr., New Fairfield, CT, 06812	Telephone: (203)746-0197	FAX:
	E-Mail Address: volga@javanet.com	Date: 09/14/00

ERIC Full Text Provided by ERIC

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or* , if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS).

Publisher/Distributor:

Address:

Price Per Copy:

Quantity Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant a reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V.WHERE TO SEND THIS FORM:

You can send this form and your document to the ERIC Clearinghouse on Languages and Linguistics, which will forward your materials to the appropriate ERIC Clearinghouse.

Acquisitions Coordinator
ERIC Clearinghouse on Languages and Linguisitics
4646 40th Street NW
Washington, DC 20016-1859

(800) 276-9834/ (202) 362-0700
e-mail: eric@cal.org

ERIC
Full Text Provided by ERIC